A Protein-Protein Docking Benchmark

Rong Chen[1], Julian Mintseris[1], Joël Janin[2] and Zhiping Weng[1,3]

[1]Bioinformatics Program, Boston University, 44 Cummington Street, Boston, MA 02215
[2]Laboratoire d'Enzymologie et de Biochimie Structurales, CNRS UPR9063, Gif-sur-Yvette, France
[3]Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, MA 02215

Corresponding Author: Zhiping Weng (zhiping@bu.edu)

ABSTRACT

We have developed a non-redundant benchmark for testing protein-protein docking algorithms. Currently it contains 59 test cases: 22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, 11 other complexes, and 7 difficult test cases. 31 of the test cases, for which the unbound structures of both the receptor and ligand are available, are classified as follows: 16 enzyme/inhibitor, 5 antibody/antigen, 5 others and 5 difficult. Such a centralized resource should benefit the docking community not only as a large curated test set but also as a common ground for comparing different algorithms. The benchmark is available at (http://zlab.bu.edu/~rong/dock/benchmark.shtml).

MOTIVATION

The protein-protein docking problem has fascinated biophysical chemists and computational biologists since the late 1970s[1, 2]. Given the 3-dimentional (3D) structures of two interacting proteins, a docking algorithm aims to determine the 3D structures of the complex. Despite the rapid accumulation of genome sequences and individual protein structures fueled by various genomics projects, experimental determination of the 3D structures of protein-protein complexes has remained difficult. Since protein complex structures can provide insights into the molecular mechanism of interaction, the protein docking problem continues to interest researches in diverse fields.

One key contributor to the difficulty of protein docking is that the interface residues of both the receptor and the ligand may undergo a conformational change upon complex formation. Frequently the conformational change is limited to side-chains; sometimes large back-bone movements are also observed. To develop predictive docking algorithms, an ideal test case must contain the unbound 3D structures of both the receptor and the ligand, in addition to the complex structure that is used only for evaluating the algorithm performance. The Protein Data Bank (PDB[3]) contains a limited number of such test cases. Therefore, it is also acceptable if one unbound structure is available, in which case the bound structure of the other molecule is used as it appears in the complex. The two classes of test cases are termed unbound/unbound and unbound/bound test cases respectively.

To facilitate the development of protein docking algorithms, we have constructed a benchmark of protein docking test cases, for which the 3D structures of the complex and one or both unbound components are available. At present, the benchmark contains all test cases used in previous docking studies, as well as additional complexes we have manually collected from the literature. Some redundant test cases have been excluded from the benchmark (see below). The benchmark should be beneficial to the entire docking field in two aspects. First, collection and curation of test cases and preparation of corresponding PDB files are tedious procedures and require biological expertise. Our effort should spare other laboratories from repeating this work. Second, the benchmark provides a common ground for comparing all docking algorithms. Different algorithms may perform better for different types of complexes, and a systematic investigation of all algorithms may reveal directions of improvement. In this sense, the benchmark complements the CAPRI effort. Even though testing on the benchmark is not "blind", it has the advantage of using a much larger set of test cases.

TREATMENT OF REDUNDANCY

In order to alleviate the bias of docking algorithms towards a limited variety of test cases, we discarded redundant complexes according to the criteria described below. A different set of criteria was used for antibody/antigen complexes since all antibodies are similar in both sequence and structure.

If two antibody/antigen systems meet *all* of the following requirements, they are redundant:

1. The antigens are the same protein.
2. Two systems share more than 80% antigen residues in the interface.

3. No major structural difference is found at the interface.

For two other systems, if the receptors and ligands meet all of the following requirements, they are redundant:

1. The E-value for sequence comparison according to BLAST[4] is smaller than 10^{-30}.
2. More than 80% of the sequences are aligned.
3. More than 60% of the sequences are identical.

The structure comparisons were performed using the K2 program[5, 6]. For redundant systems, only one is included in the benchmark, with unbound/unbound test cases taking precedence.

CURRENT STATUS OF THE BENCHMARK

Currently the benchmark (Table 1) contains 59 test cases: 22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, 11 other complexes, and 7 difficult test cases. Among them, there are 31 unbound/unbound and 28 unbound/bound test cases. Among the unbound/unbound test cases, 16 are enzyme/inhibitor, 5 antibody/antigen, 5 others and 5 difficult. To our knowledge, this is the largest collection of diverse protein-protein docking test cases. The 7 difficult test cases have significant conformational change for more than half of the interface backbone residues; they are therefore suitable for testing docking algorithms that explicitly perform backbone conformational search. The remaining test cases should be amenable to rigid-body docking algorithms with some consideration of flexibility.

We did not find any significant correlation between the resolution of the structures of the complexes or the individual protomers and the Root Mean Square Deviation (RMSD) calculated after superimposing the complexes with the corresponding protomers. This suggests that within the range we are considering (up to 3.25 Å), the resolution of the structures will not play a significant role in docking performance. The range of ΔASA (change in Accessible Surface Area on complex formation) in this dataset is 1130–4150 Å (Table 2), which encompasses the "small", "standard", and "large" interface sizes as discussed by Lo Conte et al.[7] and thus represents a good sampling of the space of protein interfaces.

The benchmark is available from the web site http://zlab.bu.edu/~rong/dock/benchmark.shtml, including the PDB ID and chain information of the corresponding proteins and cleaned-up PDB files. The most commonly used biochemical information is the Complementarity Determining Regions (CDRs) of classic antibodies. We also provide another set of PDB files for antibodies, where CDRs are indicated using sequence information, as described previously[8].

We are dedicated to maintaining and updating the benchmark. We are in the process of developing methods to automatically parse the PDB and identify additional test cases. We also welcome notifications of new test cases. With continuing growth, the benchmark should prove a valuable resource for the protein docking community. While the set described here represents the first official release of the benchmark (version 1.0), our group and others have used a version (version 0.0) to test docking algorithms. As the database grows, we will keep older releases available to the web site in order to ensure fair comparison of different algorithms on different sets.

ACKNOWLEDGEMENT

We thank Dr. Graham Smith for providing some of the test cases. JM is a DOE fellow. RC and ZW are funded by NSF grants DBI-0078194 and DBI-0133834.

REFERENCES

1. Levinthal C, Wodak SJ, Kahn P, and Dadivanian AK, Hemoglobin interaction in sickle cell fibers. I: Theoretical approaches to the molecular contacts. Proc Natl Acad Sci U S A 1975;72:1330-4.
2. Wodak SJ and Janin J, Computer analysis of protein-protein interaction. J Mol Biol 1978;124:323-42.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, and Bourne PE, The Protein Data Bank. Nucleic Acids Res 2000;28:235-42.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-402.
5. Szustakowski JD and Weng Z, Protein structure alignment using a genetic algorithm. Proteins 2000;38:428-40.
6. Szustakowski JD and Weng Z, K2: protein structure comparisons and their statistical significance, in *Evolutionary Computation in Bioinformatics*, Fogel G and Corne D, Editors. 2002, Morgan Kaufmann.
7. Lo Conte L, Chothia C, and Janin J, The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177-98.
8. Chen R and Weng Z, Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins: Structure, Function and Genetics 2002;47:281-294.
9. Hubbard SJ and Thornton JM, NACCESS. 1993, Department of Biochemistry and Molecular Biology, University College London.

Table 1: Benchmark

Complex[a]	Receptor[a]	Ligand[a]	Receptor Description	Ligand Description	RMSD[b](Å)	CA[c]	ΔASA[d](Å^2)
Enzyme-inhibitor (22)							
Unbound-unbound (16)							
1ACB(E:I)	5CHA(A)	1CSE(I)	α -chymotrypsin	Eglin C	0.7	1	1540
1AVW(A:B)	2PTN	1BA7(A)	Trypsin	Soybean Trypsin inhibitor	0.35	0	1740
1BRC(E:I)	1BRA	1AAP(A)	Trypsin	APPI	0.44	0	1320
1BRS(A:D)	1A2P(B)	1A19(A)	Barnase	Barstar	0.47	0	1560
1CGI(E:I)	1CHG	1HPT	α-chymotrypsinogen	Pancreatic secretory trypsin inhibitor	1.48	14	2050
1CHO(E:I)	5CHA(A)	2OVO	α -chymotrypsin	Ovomucoid 3rd Domain	0.59	1	1470
1CSE(E:I)	1SCD	1ACB(I)	Subtilisin Carlsberg	Eglin C	0.43	0	1490
1DFJ(I:E)	2BNH	7RSA	Ribonuclease inhibitor	Ribonuclease A	1.04	13	2580
1FSS(A:B)	2ACE(E)	1FSC	Snake Venom Acetylcholinesterase	Fasciculin II	0.75	1	1970
1MAH(A:F)	1MAA(B)	1FSC	Mouse Acetylcholinesterase	Fasciculin 2	0.6	0	2150
1TGS(Z:I)	2PTN	1HPT	Trypsinogen	Pancreatic secretory trypsin inhibitor	1.49	17	1720
1UGH(E:I)	1AKZ	1UGI(A)	Human Uracil-DNA glycosylase	Inhibitor	0.53	1	2190
2KAI(AB:I)	2PKA(XY)	6PTI	Kallikrein A	Trypsin inhibitor	0.7	2	1420
2PTC(E:I)	2PTN	6PTI	β-trypsin	Pancreatic trypsin inhibitor	0.32	0	1430
2SIC(E:I)	1SUP	3SSI	Subtilisin BPN	Subtilisin inhibitor	0.4	0	1620
2SNI(E:I)	1SUP	2CI2(I)	Subtilisin Novo	Chymotrypsin inhibitor 2	0.37	0	1630
Unbound-bound (6)							
1PPE(E:I)	2PTN	1PPE(I)	Trypsin	CMT-1	0.27	0	1690
1STF(E:I)	1PPN	1STF(I)	Papain	Stefin B	0.25	0	1790
1TAB(E:I)	2PTN	1TAB(I)	Trypsin	BBI	0.27	0	1360
1UDI(E:I)	1UDH	1UDI(I)	Virus Uracil-DNA glycosylase	Inhibitor	0.36	0	2020
2TEC(E:I)	1THM	2TEC(I)	Thermitase	Eglin C	0.19	0	1560
4HTC(LH:I)	2HNT(LCEF)	4HTC(I)	A –Thrombin	Hirudin	0.56	2	3320
Antibody-antigen (19)							
Unbound-unbound (5)							
1AHW(DE:F)	1FGN(LH)	1BOY	Antibody Fab 5G9	Tissue factor	0.71	1	1900
1BVK(DE:F)	1BVL(LH)	3LZT	Antibody Hulys11 Fv	Lysozyme	1.22	3	1400
1DQJ(AB:C)	1DQQ(LH)	3LZT	Hyhel - 63 Fab	Lysozyme	0.73	3	1760

1MLC(AB:E)	1MLB(AB)	1LZA	IgG1 D44.1 Fab fragment	Lysozyme	0.85	3	1390
1WEJ(LH:F)	1QBL(LH)	1HRC	IgG1 E8 Fab fragment	Cytochrome C	0.32	0	1180
Unbound-bound (14)							
1BQL(LH:Y)	1BQL(LH)	1DKJ	Hyhel - 5 Fab	Lysozyme	0.52	2	1630
1EO8(LH:A)	1EO8(LH)	2VIU(A)	Bh151 Fab	Influenza Virus Hemagglutinin	0.28	0	1530
1FBI(LH:X)	1FBI(LH)	1HHL	IgG1 Fab fragment	Lysozyme	0.5	0	1690
1IAI(MI:LH)	1AIF(LH)	1IAI(LH)	IgG1 Idiotypic Fab	Igg2A Anti-Idiotypic Fab	0.99	12	1890
1JHL(LH:A)	1JHL(LH)	1GHL(A)	IgG1 Fv Fragment	Lysozyme	0.26	0	1240
1KXQ(D:E)	1PIF(A)	1KXQ(E)	α-amylase	Camelid AMD9 Vhh Domain	0.43	0	2140
1KXT(A:B)	1PIF(A)	1KXT(B)	α-amylase	Camelid AMB7 Vhh Domain	0.39	0	1620
1KXV(A:C)	1PIF(A)	1KXV(C)	α-amylase	Camelid AMD10 Vhh Domain	0.24	0	1620
1MEL(B:M)	1MEL(B)	1LZA	Vh Single-Domain Antibody	Lysozyme	0.65	2	1690
1NCA(LH:N)	1NCA(LH)	7NN9	Fab NC41	Influenza Virus Neuraminidase	0.24	0	1950
1NMB(LH:N)	1NMB(LH)	7NN9	Fab NC10	Influenza Virus Neuraminidase	0.21	0	1350
1QFU(LH:A)	1QFU(LH)	2VIU(A)	Igg1-k Fab	Influenza Virus Hemagglutinin	0.27	0	1840
2JEL(LH:P)	2JEL(LH)	1POH	Jel42 Fab Fragment	A06 Phosphotransferase	0.18	0	1500
2VIR(AB:C)	2VIR(AB)	2VIU(A)	Igg1-lamda Fab	Influenza Virus Hemagglutinin	0.41	1	1260
Others (11)							
Unbound-unbound (5)							
1AVZ(B:C)	1AVV	1SHF(A)	HIV-1 NEF	FYN tyrosin kinase SH3 domain	0.73	1	1260
1L0Y(A:B)	1BEC	1B1Z(A)	T Cell Receptor β chain	Exotoxin A1	0.83	2	1130
1WQ1(G:R)	1WER	5P21	RAS activating domain	RAS	0.83	9	2910
2MTA(LH:A)	2BBK(LH)	1AAN	Methylamine dehydrogenase	Amicyanin	0.34	0	1460
2PCC(A:B)	1CCA	1YCC	Cytochrome C Peroxidase	Iso-1-Cytochrome C	0.44	1	1140
Unbound-bound (6)							
1A0O(A:B)	1CHN	1A0O(B)	Che A	Che Y	1.59	9	1130
1ATN(A:D)	1ATN(A)	3DNI	Actin	Deoxyribonuclease I	0.31	0	1770
1GLA(G:F)	1GLA(G)	1F3G	Glycerol kinase	GSF III	0.37	0	1300

1IGC(LH:A)	1IGC(LH)	1IGD	IgG1 Fab Fragment	Protein G	0.74	1	1330
1SPB(S:P)	1SUP	1SPB(P)	Subtilisin	Subtilisin prosegment	0.35	0	2230
2BTF(A:P)	2BTF(A)	1PNE	β –Actin	Profilin	0.29	0	2060

Difficult Test Cases (7)

Unbound-unbound (5)

1BTH(LH:P)	2HNT(LCEF)	6PTI	Thrombin mutant	Pancreatic trypsin inhibitor	1.91	18	2370
1FIN(A:B)	1HCL	1VIN	CDK2 cyclin-dependant kinase 2	Cyclin	4.66	59	3400
1FQ1(B:A)	1B39(A)	1FPZ(F)	CDK2	KAP	3.55	23	1830
1GOT(A:BG)	1TAG	1TBG(AE)	Transducin Gt-α, Gi-α chimera	Gt-β-γ	2.45	30	2500
1KKL(AC:H)	1JB1	1SPH(A)	HPr Kinase	Phosphocarrier Protein Hpr	2.53	28	1640

Unbound-bound (2)

1EFU*(A:B)	1D8T(A)	1EFU(B)	E. coli Ef-Tu	Efts	2.57	109	3630
3HHR*(B:A)	3HHR(B)	1HGU	Human growth hormone	Receptor	2.04	24	4150

[a] 4-letter PDB code for the crystal structures used in this study with chain IDs in parenthesis.
[b] The RMSD of the interface C_α atoms for input receptor and ligand after superposition onto the co-crystallized complex structure, calculated as in our previous work[8].
[c] Number of interface C_α atoms with RMSD larger than 2 Å between unbound and bound structures after superposition.
[d] ΔASA - change in Accessible Surface Area (ASA) upon complex formation was calculated using the program NACCESS[9].